

Running LLMs Locally With Docker Model Runner

ODSC AI West 2025

Rami Krispin, October 28th, 2025

Rami Krispin

Senior Manager - Data Science & Engineering

Author | Docker Captain  | LinkedIn Instructor

The AIOps Newsletter

Home Notes GitHub Actions Course Archive About

ISSUE 7 From Zero to a Dockerized Development Environment in Minutes with GitHub Repository Templates An effective approach for setting up a development environment SEP 21 · RAMI KRISPIN

Running OpenAI GPT OSS Locally with Docker Model Runner and R This is the fourth tutorial in the Docker Model Runner sequence AUG 19 · RAMI KRISPIN

Docker Model Runner - Pull LLMs from Hugging Face This is the third tutorial in the Docker Model Runner sequence AUG 8 · RAMI KRISPIN

Running LLMs Locally with Docker Model Runner and Python This is the second tutorial on the Docker Model Runner sequence JUL 29 · RAMI KRISPIN

Most Popular

Getting Started with Docker Running OpenAI GPT OSS Introduction to Docker From Zero to a Dockerized

VIEW ALL



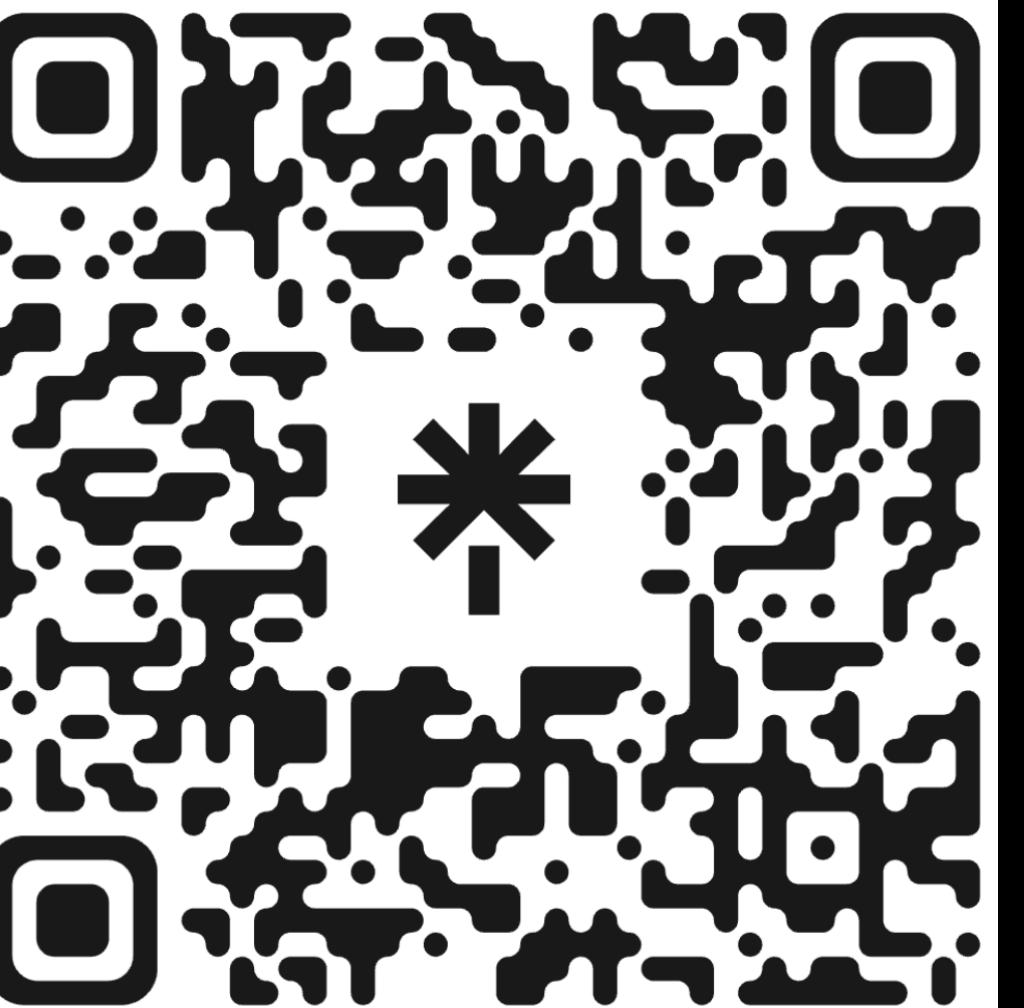
Running LLMs with Docker Desktop
The coming release of Docker Desktop from Docker, Inc is going t...
3 min read



The skforecast Project, AI Engineering and New Learnin...
Happy Saturday! A quick update - the newsletter is moving from...
3 min read



The Optuna Project, Advanced Topics in Cryptography, Beyo...
This week's agenda: Open Source of the Week - the Optuna project New...
3 min read



Agenda

- Introduction to DMR
- Setup
- Pull LLMs from Docker Hub & Hugging Face
- Running LLMs via the CLI
- Running LLMs with Python
- OpenCode

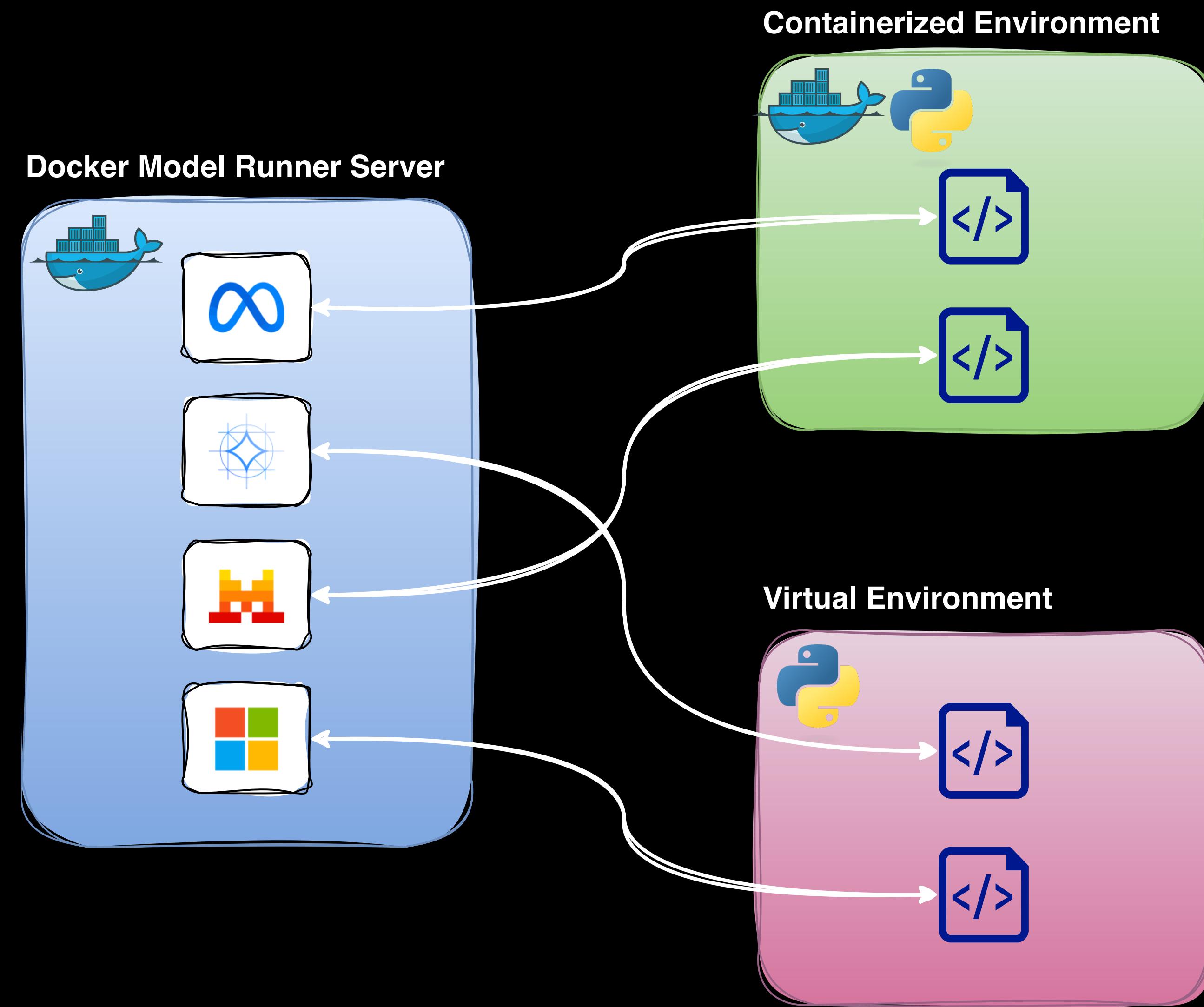
Poll

- Docker
- Python
- OpenAI Python SDK

[https://github.com/RamiKrispin/
odsc-ai-2025-dmr-workshop](https://github.com/RamiKrispin/odsc-ai-2025-dmr-workshop)

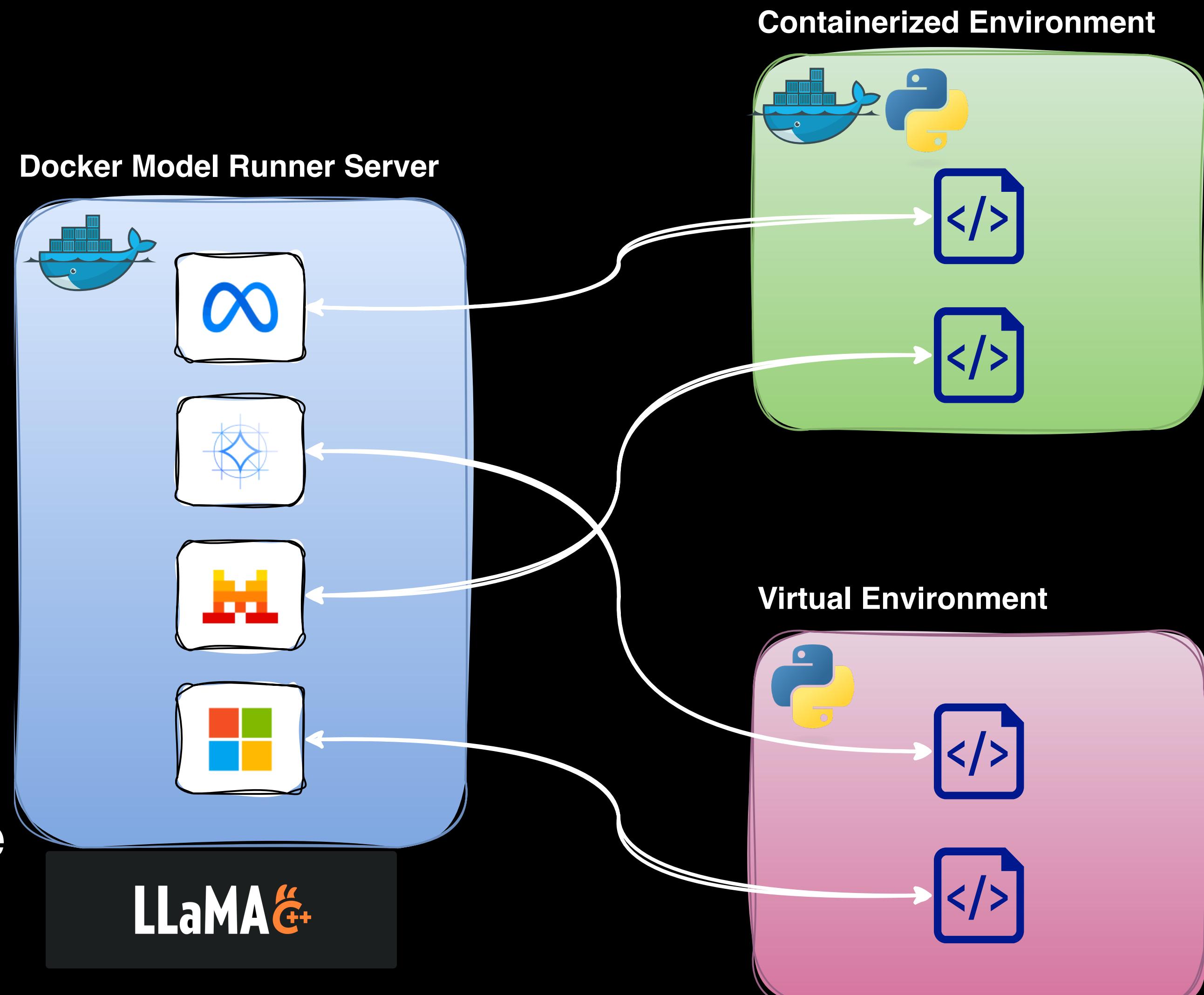
DMR Doesn't Require Prior Docker Knowledge

Introduction to DMR



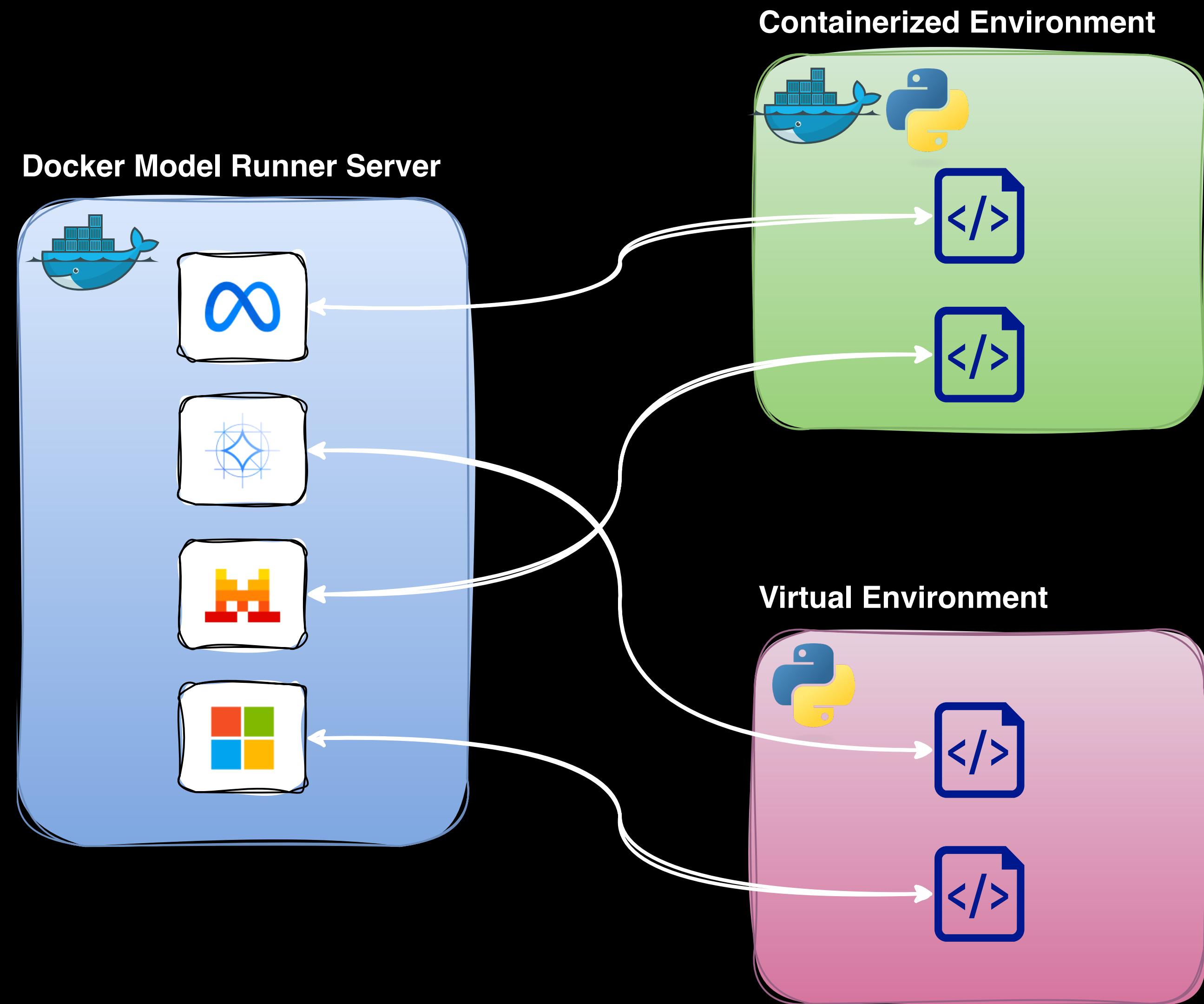
Introduction to DMR

- Docker Desktop feature
- Based on Llama.cpp
- Easy to setup and use
- Open Container Initiative
- Supports GGUF format
- OpenAI API SDK compatible



Applications

- Local development of AI applications
- Running local AI applications (local AI agents, code generation, etc.)
- Privacy



DMR Backend



General Requirements

- Docker Engine (Linux)
- Docker Desktop 4.40 (Mac) or 4.41 (Windows)
- Docker Hub account
- Hugging Face API key (optional)
- OpenCode (optional)
- Local computational resources

Setup

Setup

The screenshot shows the Docker Desktop landing page. At the top, there is a banner with the text "New Docker + E2B. A new partnership bringing trust to AI development. Learn more." followed by a close button. Below the banner is a dark blue header bar with the Docker logo, navigation links for AI, Products, Developers, Pricing, Support, Blog, and Company, and buttons for Sign In and Get Started. The main content area features the Docker Desktop logo and the text "The #1 containerization software for developers and teams". Below this, a subtitle reads "Streamline development with Docker Desktop's powerful container tools." At the bottom, there are two buttons: "Choose plan" and "Download Docker Desktop", with a pink arrow pointing to the "Download Docker Desktop" button. The footer shows the Docker Desktop interface with sections for Containers, Images, Volumes, Builds, and Dev Environments, along with real-time usage statistics for CPU and memory.

↑ New Docker + E2B. A new partnership bringing trust to AI development. Learn more. →

Docs Get Support Contact Sales

docker

AI Products Developers Pricing Support Blog Company

Sign In Get Started

docker desktop

The #1 containerization software for developers and teams

Streamline development with Docker Desktop's powerful container tools.

Choose plan Download Docker Desktop

Containers Images Volumes Builds Dev Environments

Containers Give feedback View all your running containers and applications. Learn more

Container CPU usage 0.01% / 1200% (12 CPUs allocated)

Container memory usage 249.6MB / 7.47 GB

Search Only show running containers

Name Container ID Image Port(s) Last started CPU % Actions

ODSC AI WEST 2025

Setup



Screenshot of the Docker Hub homepage:

Docker Hardened Images - Secure & Compliant
Enterprise-grade Docker images with built-in security, compliance, and continuous updates. Minimize vulnerabilities and deploy with confidence.

[Visit catalog now](#)

Generative AI

- AI Models
- MCP Servers

Trusted content

- Docker Hardened Images
- Docker Official Images
- Verified Publisher
- Sponsored OSS

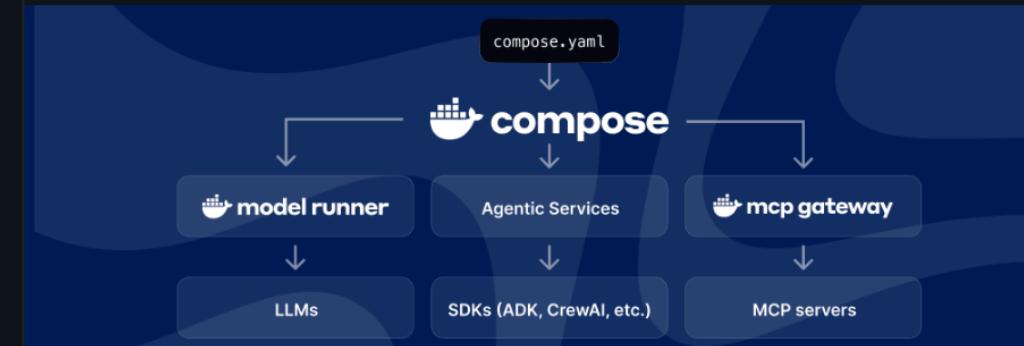
Categories

- Networking
- Security
- Languages & frameworks

Spotlight

AI MEETS COMPOSE
Build AI Agents Faster with Docker Compose

Use the workflow you know to develop and deploy across local, cloud, and multi-cloud environments with Docker Compose.



MCP
E2B + Docker: Trusted AI

Every E2B sandbox includes direct access to Docker's MCP Catalog, a collection of 200+ tools such as GitHub, Perplexity, Browserbase, and ElevenLabs, all enabled by the Docker MCP Gateway.



SOFTWARE SUPPLY CHAIN
Secure Your Supply Chain with Docker Hardened Images

Use Docker's enterprise-grade base images: secure, stable, and backed by SLAs for Ubuntu, Debian, Java, and more. Regularly scanned and maintained with CVE remediation and long-term support.



Setup

The screenshot shows the Hugging Face homepage with a dark theme. On the right side, a user profile dropdown menu is open for the user "RamiKrispin". The menu includes options like "Profile", "Notifications", "Inbox (0)", "Trending", "Create organization", "Usage Quota", "Private Storage", "Zero GPU", "Inference Usage", "Get Hugging Face PRO", "Settings", and "Access Tokens". A large yellow arrow points to the "Access Tokens" link. The main content area displays the user's activity feed, which includes posts about updating spaces and liking models.

Hugging Face Search models, datasets, users...

Models Datasets Spaces Community Docs Enterprise Pricing

RamiKrispin

- + New
- Profile
- Inbox (0)
- Settings
- Billing
- Get PRO

Organizations

- Create New

Resources

- Getting Started
- Documentation
- Forum
- Tasks
- Learn

Light theme

My activity

- All Models Datasets Spaces Papers Collections Community Posts Upvotes Likes Articles

Updated 2 Spaces over 1 year ago

- My Streamlit
- Shiny for Python template

Updated 2 Spaces almost 2 years ago

- Docker Poc
- Shiny for R template

Liked a model about 2 years ago

- Deci/DeciLM-6b-instruct

Text Generation · 6B · Updated Feb 15, 2024 · 12 · 132

Trending last 7 d

- deepseek-ai/De
- PaddlePaddle/P
- tencent/Hunyu
- krea/krea-real
- HuggingFaceFW/
- DeepSeek OCR Demo
- veo3.1-fast
- DeepSite v3
- Wan2.2 Animate
- karpathy/fineweb-edu-100b-shuffle
- nick007x/github-code-2025

Profile RamiKrispin

- Notifications
- Inbox (0)

+ New Model
+ New Dataset
+ New Space
+ New Collection

Create organization

Usage Quota

Private Storage 0 GB/100 GB
Zero GPU 0/4 min
Inference Usage \$0.00 / \$0.10

Get Hugging Face PRO →

Settings

Access Tokens

Billing

Changelog • 3

Sign Out

2.07k

38.2k · 109

147M · 11.3k · 78

ODSC AI WEST 2025

Setup

The easiest way to install OpenCode is through the install script.

```
curl -fsSL https://opencode.ai/install | bash
```

You can also install it with the following commands:

Using Node.js

NPM BUN PNPM YARN

```
npm install -g opencode-ai
```

Using Homebrew on macOS and Linux

```
brew install sst/tap/opencode
```

Using Paru on Arch Linux

```
paru -S opencode-bin
```

Setup

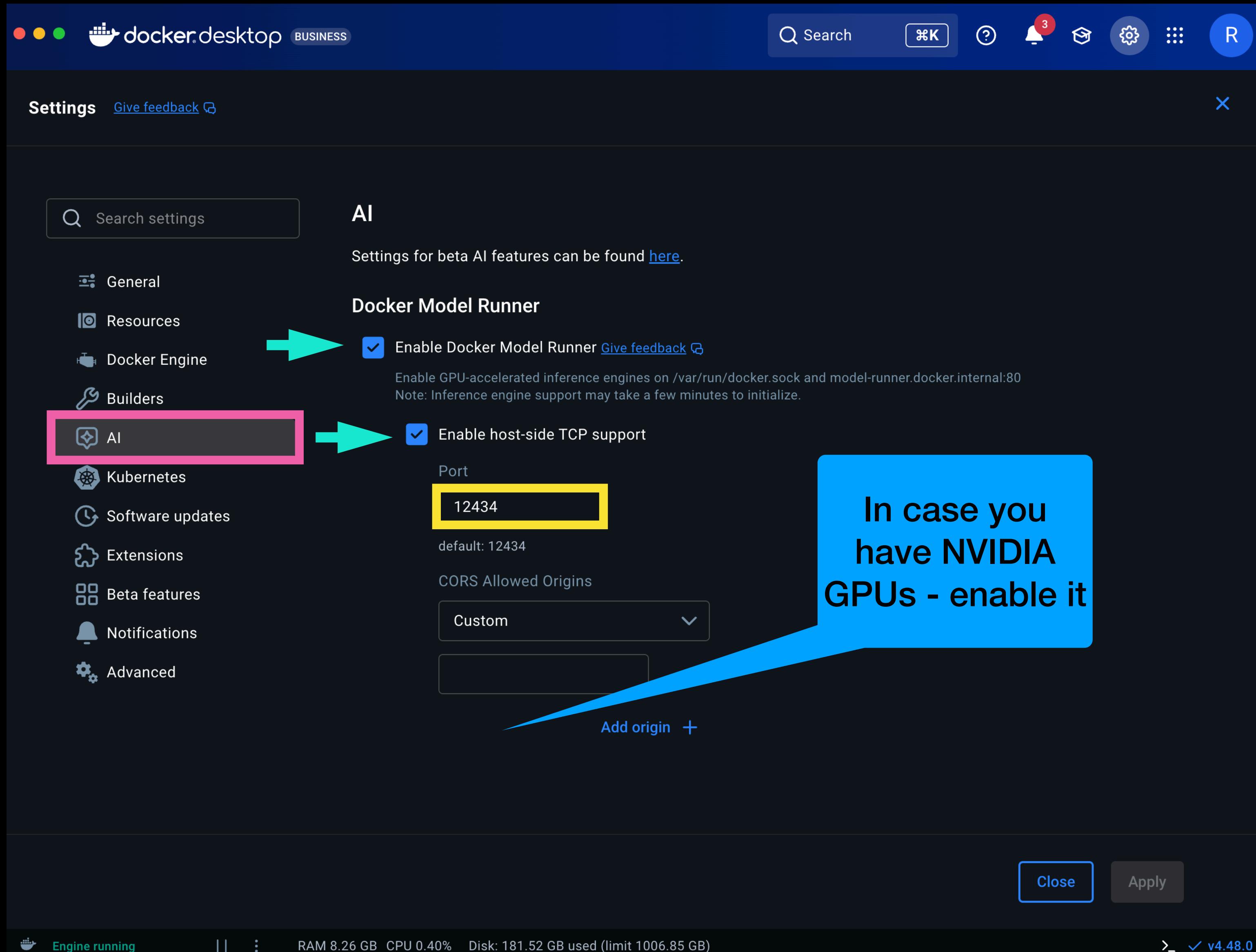
```
ramikrispin ~ 17:06 docker login
Authenticating with existing credentials... [Username: rkrispin]

i Info → To login with a different account, run 'docker logout' followed by 'docker login'

Login Succeeded

ramikrispin ~ 17:07 hf auth login --token $HF_TOKEN
The token has not been saved to the git credentials helper. Pass `add_to_git_credential=True` in this
function directly or `--add-to-git-credential` if using via `hf` CLI if you want to set the git credential
as well.
Token is valid (permission: fineGrained).
The token `DMR Workshop` has been saved to /Users/ramikrispin/.cache/huggingface/stored_tokens
Your token has been saved to /Users/ramikrispin/.cache/huggingface/token
Login successful.
Note: Environment variable `HF_TOKEN` is set and is the current active token independently from the tok
en you've just configured.
```

Setup



Models

✖ New Docker + E2B. A new partnership bringing trust to AI development. Learn more. → X

hub Search Docker Hub 3K ⓘ ⚡ ⚡ Sign in Sign up

Docker Verified Publisher

Verified Publisher Docker San Francisco, CA, USA https://www.docker.com

Repositories

Search by repository name Displaying 1 to 30 of 32 repositories

MODEL	NAME	OWNER	PULLS	STARS	LAST UPDATED
ai/granite-docling	ai/granite-docling	IBM Docker	5.7K	1	17 days
Granite Docling is a multimodal model for efficient document conversion.					
Pulls	5.7K				
Stars	1				
Last Updated	17 days				
ai/granite-4.0-h-small	ai/granite-4.0-h-small	IBM Docker	3.1K	1	25 days
32B long-context instruct model with RL alignment, IF, tool use, and enterprise optimization.					
Pulls	3.1K				
Stars	1				
Last Updated	25 days				
ai/moondream2	ai/moondream2	IBM Docker	2.9K	0	17 days
An open-source visual language model that interprets images via text prompts, fast and powerful.					
Pulls	2.9K				
Stars	0				
Last Updated	17 days				
ai/smolvlm	ai/smolvlm	IBM Docker	4.1K	2	18 days
SmolVLM: lightweight multimodal model for video, image, and text analysis, optimized for devices.					
Pulls	4.1K				
Stars	2				
Last Updated	18 days				
ai/granite-4.0-h-tiny	ai/granite-4.0-h-tiny	IBM Docker	3.1K	2	25 days
7B long-context instruct model with RL alignment, IF, tool use, and enterprise optimization.					
Pulls	3.1K				
Stars	2				
Last Updated	25 days				
ai/granite-4.0-h-micro	ai/granite-4.0-h-micro	IBM Docker	2.0K	2	25 days
3B long-context instruct model with RL alignment, IF, tool calling, and enterprise readiness.					
Pulls	2.0K				
Stars	2				
Last Updated	25 days				

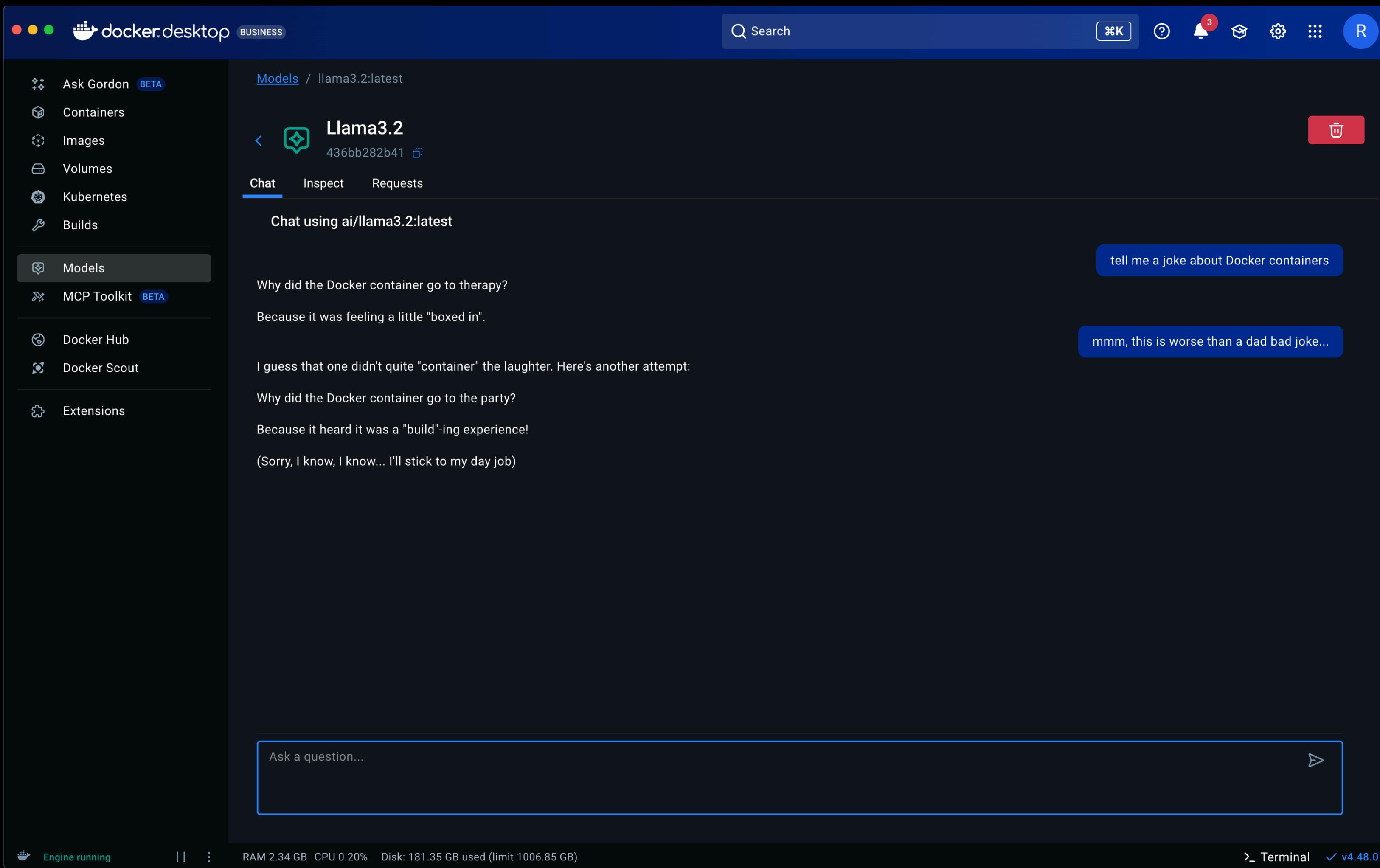
Models

The screenshot shows the Docker Desktop application interface. The left sidebar has a dark theme with white icons and text. The 'Models' option is selected, highlighted with a grey background. Other options include 'Ask Gordon (BETA)', 'Containers', 'Images', 'Volumes', 'Kubernetes', 'Builds', 'MCP Toolkit (BETA)', 'Docker Hub', 'Docker Scout', and 'Extensions'. The main area is titled 'Models' and shows a grid of nine Docker Hub model cards. Each card includes the model name, a 'Pull' button, a brief description, download statistics (e.g., 5.7K), and a star rating. A navigation bar at the bottom shows page 1 of 2.

Model Name	Description	Downloads	Stars
granite-docling	Granite Docling is a multimodal model for efficient document conversion.	5.7K	1
moondream2	An open-source visual language model that interprets images via text prompts, fast and powerful.	2.9K	0
smolvlm	SmolVLM: lightweight multimodal model for video, image, and text analysis, optimized for devices.	4.1K	2
granite-4.0-h-small	32B long-context instruct model with RL alignment, IF, tool use, and enterprise optimization.	3.1K	1
granite-4.0-h-tiny	7B long-context instruct model with RL alignment, IF, tool use, and enterprise optimization.	3.1K	2
granite-4.0-h-micro	3B long-context instruct model with RL alignment, IF, tool calling, and enterprise readiness.	2.0K	2
granite-4.0-micro	3B long-context instruct model with RL alignment, IF, tool use, and enterprise optimization.	1.4K	0
devstral-small	Agentic coding LLM (24B) fine-tuned from Mistral-Small-3.1 with a 128K context window	3.5K	2
magistral-small-3.2	24B multimodal instruction model by Mistral AI, tuned for accuracy, tool use & fewer repeats	7.0K	1

Hello World!

Running LLMs via Docker Desktop



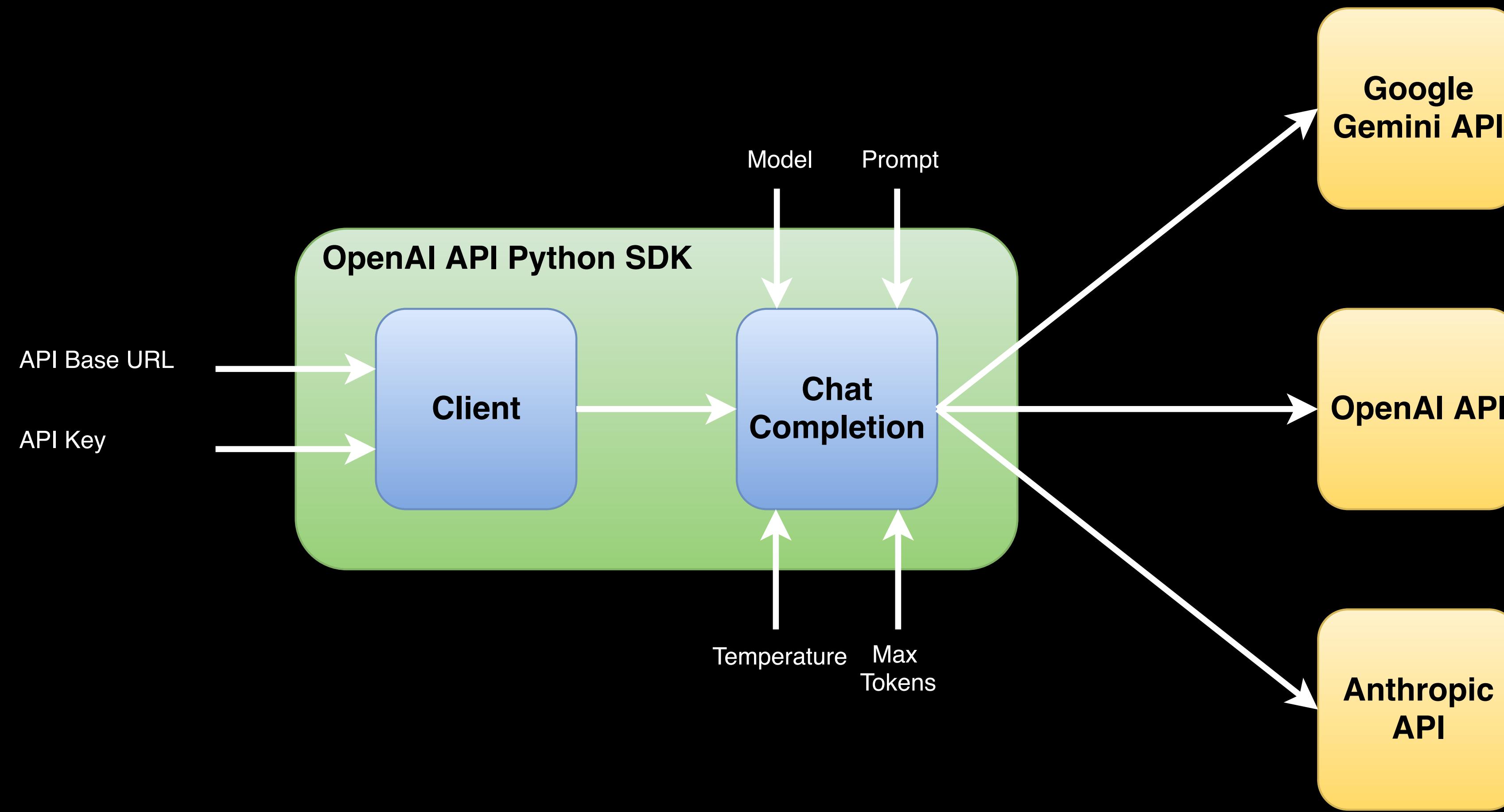
Running LLMs via the CLI

- CLI core commands
- Running LLMs
- Pull models from Docker Hub
- Pull LLMs from Hugging Face
- Update the context window

Running LLMs with Python

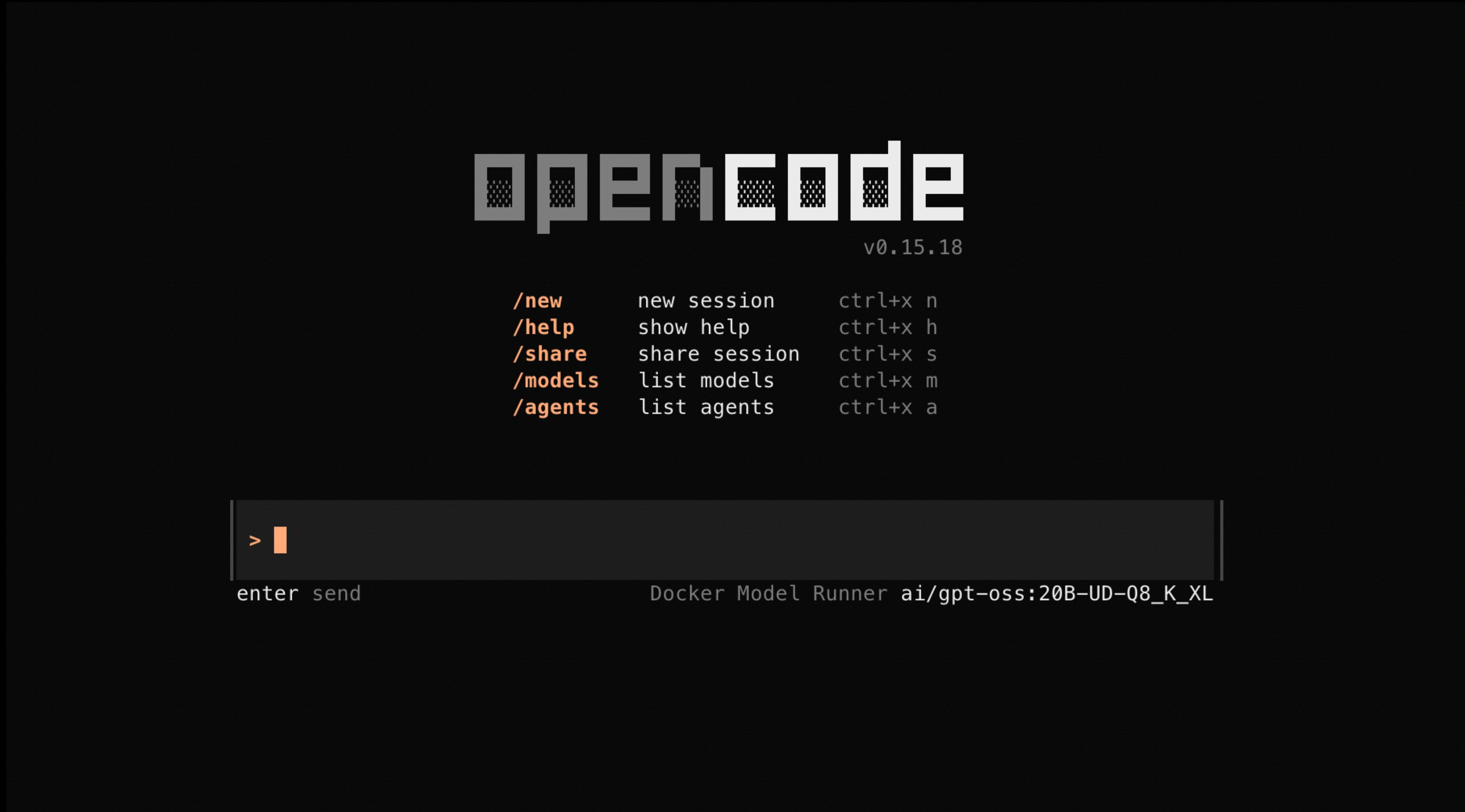
- OpenAI API SDK
- Calling LLMs from Virtual Environment vs. Container
- LangChain

OpenAI API SDK



Demo

OpenCode



OpenCode

```
{  
  "$schema": "https://opencode.ai/config.json",  
  "provider": {  
    "dmr": {  
      "npm": "@ai-sdk/openai-compatible",  
      "name": "Docker Model Runner",  
      "options": {  
        // Case using inside a container:  
        // "baseURL": "http://model-runner.docker.internal/engines/v1",  
        // Otherwise use the following base URL:  
        "baseURL": "http://localhost:12434/engines/v1",  
        "apiKey": "docker"  
      },  
      "models": {  
        "ai/gpt-oss:20B-UD-Q8_K_XL": {  
          "name": "ai/gpt-oss:20B-UD-Q8_K_XL"  
        },  
        "ai/llama3.2:3B-Q4_0": {  
          "name": "ai/llama3.2:3B-Q4_0"  
        },  
        "ai/devstral-small:24B": {  
          "name": "ai/devstral-small:24B"  
        }  
      }  
    }  
  }  
}
```

Resources

- DMR documentation - <https://docs.docker.com/ai/model-runner/>
- Docker podcast - <https://www.youtube.com/watch?v=9XryO1Oxb4A>
- Workshop repo -
- The AIOps Newsletter - <https://theaiops.substack.com/>